FEBS Letters 580 (2006) 3895–3899

Temperature adaptation of synonymous codon usage in different functional categories of genes: A comparative study between homologous genes of *Methanococcus jannaschii* and *Methanococcus maripaludis*

Surajit Basak, Tapash Chandra Ghosh*

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

Received 13 April 2006; revised 29 May 2006; accepted 6 June 2006

Available online 16 June 2006

Edited by Takashi Gojobori

Abstract **Synonymous codon usage of homologous sequences between *Methanococcus jannaschii* and *Methanococcus maripaludis* have been analyzed in three broad functional categories of genes namely: (i) information storage and processing; (ii) metabolism; and (iii) cellular processes and signaling. Average values of synonymous nucleotide substitutions per synonymous site are significantly lower for information processing genes compared to either metabolic or cellular processing genes. These results suggests that synonymous codon usage has been subject to greater constraint in the information storage and processing group of genes compared to other functional categories of genes. For metabolic and cellular processing genes, correspondence analysis based on relative synonymous codon usage (RSCU) values separates the genes along the first major axes according to the genome type; while in the information processing group, genes are separated along the second major axes according to the genome type. Further study on synonymous substitution rate for information processing genes shows a stronger selective constraint on synonymous codon usage of six amino acids, G, A, R, P, Y, F. Randomization of the original transcript of *M. jannaschii* for information processing genes suggests that variation in selective constraint between synonymous codon usage is related to the potential formation of mRNA secondary structures which contribute to the folding stability.**

Keywords: Optimal growth temperature; Gene functional categories; Synonymous codon usage; Synonymous substitution; Non-synonymous substitution; Free folding energy

1. Introduction

Non-random usages of synonymous codons both within and between organisms are well documented in the literature [1–3]. Biased codon usage may arise from various factors. It has been reported that mutational bias and/or selective forces are the main driving force for the codon usage variation among the genes in different organisms [4–9]. In some unicellular organisms, it has been demonstrated that translational selection is the main factor in driving the codon usage variation among the genes in these organisms as it has been observed that preferred codons in highly expressed genes are recognized by most abundant tRNAs present in the cell [10–12]. Sharp and Li [13] have demonstrated that synonymous substitution rate is significantly lower in highly biased genes than in those genes having lower codon bias. Some recent results also suggest that difference in codon usage is related to gene function [14,15].

Recently it has been reported that the thermophilic organisms have a distinct pattern of synonymous codon usage compared to mesophilic organisms [16] and subsequently it has been shown that the difference in synonymous codon usage between thermophilic and mesophilic organisms is independent of protein secondary structures [9]. However, variation of synonymous codon usage between thermophilic and mesophilic prokaryotes among different functional categories of genes have not yet been studied.

Here, in this study, we have culled 369 homologous gene pairs from two archaea belong to same genus, *Methanococcus jannaschii* and *Methanococcus maripaludis*, having similar genomic G+C level but with different optimal growth temperatures. We estimated the number of synonymous nucleotide substitutions per synonymous site (dS) and the number of non-synonymous nucleotide substitutions per non-synonymous site (dN) in three broad functional categories of genes. We observed a significant lowering of dS for information processing genes of both the organisms, which suggests that synonymous codon usage has been subject to greater constraint in this group of genes compared to other functional categories. Correspondence analysis based on RSCU values further suggests that evolutionary constraints act differently in the three functional categories of genes. It has also been demonstrated that selection pressure also varies between different synonymous group of amino acids and the greater constraint in the selection of synonymous codons of GC-rich amino acids and aromatic amino acids for information processing genes favors the potential formation of mRNA secondary structures which contribute to folding stability.

2. Materials and methods

The complete genome sequences of *M. jannaschii* and *M. maripaludis* have been downloaded from ftp.ncbi.nlm.nih.gov/genbank/genomes. Although, the growth temperatures of these two organisms are widely different but their genomic G+C contents are nearly identical. These

*Corresponding author. Fax: +91 33 2355 3886.
E-mail addresses: tapash@bic.boseinst.ernet.in, tapash@boseinst.ernet.in (T.C. Ghosh).

0014-5793/$32.00
doi:10.1016/j.febslet.2006.06.014

two organisms were used previously by many authors for studying the temperature adaptation of proteins [17,18]. *M. maripaludis* has a genomic G+C content of 33.1% [19] and an optimal growth temperature of 35 °C [18] whereas, *M. jannaschii* has an optimal growth temperature of 85 °C [20] and a G+C content of 31.84% ([17], www.kazusa.or.jp/codon). Our own program developed in C was used to retrieve the coding sequences from the complete genome.

Homologous sequences between *M. jannaschii* and *M. maripaludis* were identified by gapped BLASTP program [21] using cutoff of $E = 1.0 \times 10^{-3}$. Hits less than 50% identity were removed from the dataset. All hypothetical coding sequences, as well as genes having less than 100 codons were ignored. Gene pairs having size difference lower than or equal to 30 codons were retained. Finally 369 gene pairs were selected for data analysis. All the gene pairs have been classified in three broad functional categories of genes namely: (i) information storage and processing; (ii) metabolism; and (iii) cellular processes and signaling as described in COG database [22]. As a result, information storage category consists of 108 gene pairs, metabolic category consists of 228 gene pairs and cellular processes and signaling category consists of 33 gene pairs. Pair-wise synonymous (d*S*) and non-synonymous (d*N*) distance between the homologous genes of *M. jannaschii* and *M. maripaludis* was calculated by using the method of Yang and Neilsen [23]. Synonymous codon usage bias was measured by calculating the 'effective number of codons used in a gene' (N_c) [24,25]. The values of N_c range from 20 (when one codon is used per amino acid) to 61 (when all the codons are used with equal probability). Correspondence analysis [26] available in CodonW 1.4.2 (J. Peden, 2000; http://www.molbiol.ox.ac.uk/cu/) was used to investigate the major trend in relative synonymous codon usage variation among the genes. The Student's *t* test and permutation test available in *R*-statistical package were used to evaluate the significance of pairwise differences in synonymous and non-synonymous substitutions in different functional categories of genes. For each native mRNA sequence, 50 random sequences were generated using the randomization protocol, CodonShuffle [27], which randomly permutes synonymous codon in codon degenerate family preserving the exact count of each codon and order of encoded amino acids as in the original transcript. The Zipfold program was used to predict free-folding energies for each native mRNA sequence and corresponding shuffled sequence available at http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form4.cgi.

3. Results and discussion

We estimated the number of synonymous nucleotide substitutions per synonymous site (d*S*) and the number of non-synonymous nucleotide substitutions per non-synonymous site (d*N*) in three broad functional categories of genes (Information storage and processing, Metabolism, Cellular processes and signaling) of *M. jannaschii* and *M. maripaludis*. Student's *t* test was used to determine the *P*-values for the difference in average values of either d*N* or d*S* between any two of the three functional categories of genes. Table 1 shows the results. It is evident from Table 1 that irrespective of the functional categories, d*N* is always significantly lower than d*S* ($P < 0.001$ in all the functional categories). On the other hand, d*N* does not vary significantly among the functional categories but a significant lowering of d*S* has been observed for information processing genes compared to metabolic genes ($P < 0.001$) and cellular processes and signaling ($P < 0.05$) (Table 1). These results were further confirmed by permutation test, where d*S* values of information processing genes is significantly lower than the d*S* values of metabolic genes ($P < 0.05$) and cellular processes and signaling genes ($P < 0.001$). On the basis of this result, we should expect more highly biased codon usage in information processing genes for both the species because; lower value of d*S* for information processing genes indicates greater constraints on codon usage on this group of genes compared to either metabolic or cellular processes and signaling genes. In order to validate the above hypothesis, we calculated the "effective number of codons used by a gene" (N_c) for two organisms in three functional categories of genes (Table 2). The average value of N_c for information processing genes of *M. jannaschii* is significantly lower than the average values of N_c of metabolic genes ($P < 0.001$) and cellular processes and signaling genes ($P < 0.001$). Similarly, the average value of N_c for information processing genes of *M. maripaludis* is significantly lower than the average values of N_c of metabolic genes ($P < 0.05$) and cellular processes and signaling genes ($P < 0.01$). These results further support our notion that codon usage bias for both the organisms is stronger in information processing genes compared to other two functional categories of genes.

In order to see how the optimal growth temperatures impose selective constraints on codon usage of different functional categories of genes, we have performed correspondence analysis in three broad functional categories of genes. Since codon usage by its very nature is multivariate, one of the most popular multivariate methods for studying codon usage variation is correspondence analysis [26]. Correspondence analysis identifies the major trends in the variation of the synonymous codon usage data and distributes genes along continuous axes in accordance with these trends. Correspondence analysis on relative synonymous codon usage (RSCU) of information processing genes detected two major trends on first and second axis of inertia. The first axis accounted for 15.67% of the total variation and second axis accounted for 13.77% of the total variation. When correspondence analysis was performed on RSCU for metabolic genes, first major axes exhibited maximum variation of 17.75%, and second major axes accounted for 9.86% of the total variation. Correspondence analysis on RSCU of cellular genes detected one major trend on the first major axis and accounted for 18.56% of the total variation, second major axis accounted for 10.48% of the total variation. In all the three functional categories of genes, none of the other axes accounted for more than 7.20% of the total variation. The positions of the genes along the first and second major axes produced by correspondence analysis on RSCU values for three different functional categories are shown in Figs. 1–3. From these three figures it is evident that, for metabolic and cellular processing genes, genes are separated along the first major axis according to the genome type; while in the information processing group, genes are separated along the second major axis according to the genome type. Non-parametric Spearman correlation coefficients (ρ) have been calculated be-

Table 1
Average d*N*, d*S* values for three functional categories of genes

	Information-processing	Metabolic	Cellular
d*N*	0.265	0.264	0.299
d*S*	3.227	3.650	3.645

Table 2
Average values of N_c for three functional categories of genes

	Information-processing	Metabolic	Cellular
M. jannaschii	36.595	37.466	37.977
M. maripaludis	38.237	40.842	41.710

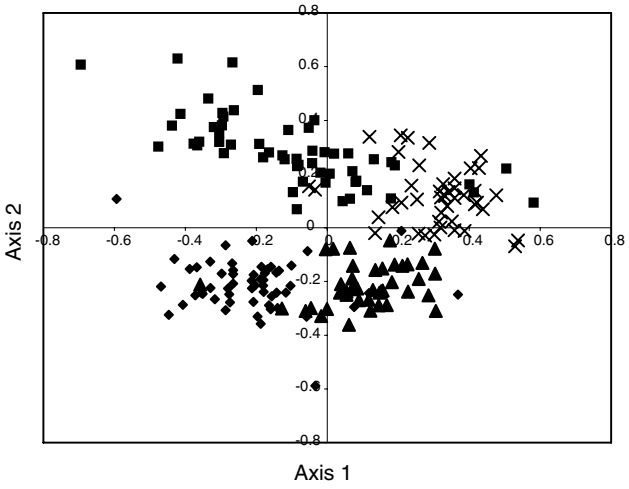

Fig. 1. Positions of the genes along the first two major axes in the correspondence analysis based on RSCU values of information storage and processing genes. Square boxes and '×' represent putatively low- and high-expression genes of *M. maripaludis* respectively. Diamonds and triangle represent putatively low- and high-expression genes of *M. jannaschii*, respectively. Square represents *M. jannaschii* genes and triangle represents *M. maripaludis* genes.

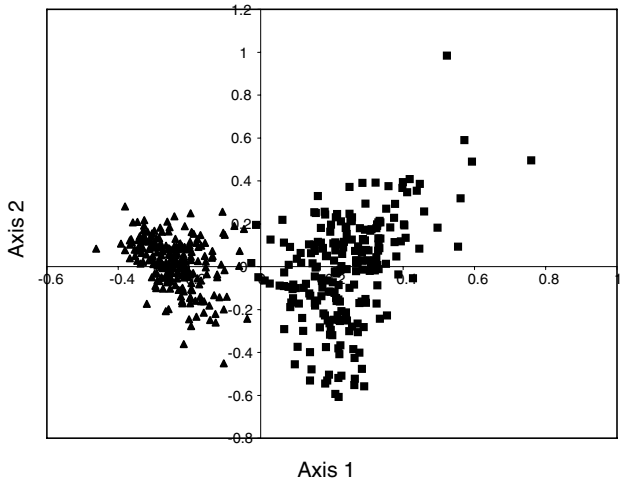

Fig. 2. Positions of the genes along the first two major axes in the correspondence analysis based on RSCU values of metabolic genes. Square represents *M. jannaschii* genes and triangle represents *M. maripaludis* genes.

tween the position of the information processing genes along the first major axis and GC content at different codon positions. The positions of the genes along the first axis in Fig. 1 correlates positively with GC content at first ($\rho = 0.300$, $P < 0.01$) and second codon positions ($\rho = 0.723$, $P < 0.01$) and also GC content of the genes ($\rho = 0.667$, $P < 0.01$). But no significant correlation has been observed between the positions of the genes along the first axis in Fig. 1 with the GC content at third codon positions. Axis 1 correlates negatively with aromaticity levels of each gene ($\rho = -0.519$, $P < 0.01$). The first axis also clearly distinguishes ribosomal protein genes (putatively highly expressed) from the regulatory genes (putatively lowly expressed) of both *M. jannaschii* and *M. maripaludis* (Fig. 1) indicating that the information processing genes may have a greater range of expression levels than their

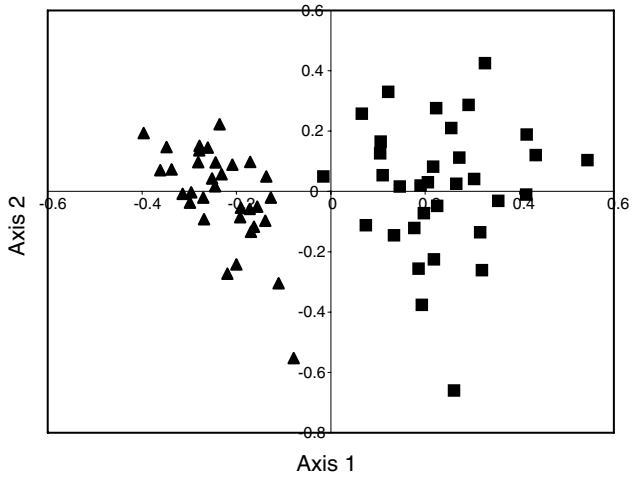

Fig. 3. Positions of the genes along the first two major axes in the correspondence analysis based on RSCU values of cellular processing genes. Square represents *M. jannaschii* genes and triangle represents *M. maripaludis* genes.

cellular and metabolic genes. If we look into the corresponding distribution of codons for three broad functional categories of gene (Supplementary tables 1–3), it is evident that the major contributors to this pattern for all the functional categories are the arginine codons (CGN and AGR), although many other codon groups also contribute to the separation between the *M. jannaschii* and *M. maripaludis*.

Interestingly, when correspondence analysis of information processing genes were performed on the basis of RSCU values after excluding: (i) codons representing aromatic amino acids namely, Phe and Tyr; and (ii) all the codons from Arg, Pro, Ala and Gly (all are GC-rich amino acids), genes have been separated along the first major axis on the basis of the genome type (data not shown). These results lead us to further investigate whether constraint on synonymous codon usage vary among different synonymous codon groups and more precisely to examine the variation of synonymous substitution rate in GC-rich and aromatic amino acids from other amino acids in the information processing group of genes. For these we have calculated synonymous substitution rate for two codon groups; one containing all the GC-rich amino acids and aromatic amino acids and another containing remaining amino acids (Table 3). A significant reduction in synonymous substitution rate has been observed ($P < 0.001$) for the codon group containing the GC-rich amino acids and aromatic amino acids with respect to other synonymous codon groups. The reduction in synonymous

Table 3
Average synonymous substitution rate (d*S*) of native and shuffled sequences of three functional categories of genes

	d*S* (native)	d*S* (shuffled)
Information-processing (Group 1)	2.6662	3.1363
Information-processing (Group 2)	2.0663	1.9954
Metabolic (Group 1)	3.2322	3.2926
Metabolic (Group 2)	1.7049	1.7867
Cellular (Group 1)	3.2330	3.3112
Cellular (Group 2)	2.0260	2.0083

Group 1 consists of amino acids L, S, T, V, K, N, Q, H, E, D, C, I, M, W. Group 2 consists of G, A, R, P, Y, F.

substitution rate reflects a stronger selective constraint on synonymous codon usage of GC-rich amino acids as well as aromatic amino acids. A possible explanation for this observation might be the presence of some constraints upon mRNA secondary structure [28]. Eyre-Walker and Bulmer [28] argued that reduction in synonymous substitution rate at the start of *Escherichia coli* genes is the result of selection to avoid mRNA secondary structure. To investigate if some substitutions are really under selection to diminish mRNA stability, the genes from three functional categories of *M. jannaschii* have been randomized using the randomization protocol, CodonShuffle [27]. For each native sequence, average value of the free folding energy of all the shuffled sequences has been calculated and a significantly higher ($P < 0.05$) free folding energy has been observed for shuffled sequence compared to native sequence only in the information processing group of genes, indicating additional selection pressure of mRNA secondary structure forming potential on codon usage in this group of genes. This is consistent with a significantly higher average synonymous substitution rate ($P < 0.001$) for shuffled sequences than native sequence for the group of amino acids excluding G, A, R, P, Y, F in information processing genes (Table 3). When the same analysis was performed on the genes of other functional categories, no significant difference in synonymous substitution rate has been observed between native sequence and shuffled sequences in either metabolic or cellular processing group of genes in two amino acids groups; one consists of G, A, R, P, Y, F and the other consists of L, S, T, V, K, N, Q, H, E, D, C, I, M, W (Table 3).

Considering the above results it is reasonable to conclude that the influence of optimal growth temperature is not equally effective on the synonymous codon usage in the three major groups of gene. While synonymous codon usage of metabolic and cellular processing groups behave similarly towards the higher growth temperature, synonymous codon usage of information processing group has been shaped by the greater variation in the expression level of the genes and a smaller fraction of the total variation in codon usage might be due to the variation in synonymous codon usage between the species (Fig. 1). Because of the closely relatedness of *M. jannaschii* and *M. maripaludis*, evolutionary significance of the present study might be drawn considering the fact that, switch to mesophile is an ongoing process in *M. maripaludis*, that is not yet complete. Possibly the switching process of information processing genes are proceeding in the same direction as the other genes, but at a slower rate compared to other genes. For information processing genes, the greater constraint on synonymous codon usage of GC-rich amino acids and aromatic amino acids is actually the causative factor to maintain the stability of mRNA secondary structure. Although Lynn et al. [16] demonstrated that the optimal growth temperature affects synonymous codon usage of thermophilic bacteria; the present study reveals that the strength of this selection pressure varies depending upon the functionality of the genes. Study on synonymous substitution rate clearly shows that selection pressure also varies between different synonymous group of amino acids and the greater constraint in the selection of synonymous codons of GC-rich amino acids and aromatic amino acids for information processing genes favors the potential formation of mRNA secondary structures which contribute to folding stability.

To conclude, while studying the temperature adaptation of various macromolecules at different growth temperature, the environments of any mesophile/thermophile pair often differ in other environmental variables also. Any of these variables other than optimal growth temperature might cause a difference in various molecular levels. Keeping these facts in mind, we have chosen two organisms in such a way that belong to the same genus where phylogenetic relationships are expected to be fairly accurate. These two bacteria can be used for the comparative evolutionary studies on synonymous codon usage as they have similar genomic G+C level, thus mutational bias is less operative in dictating the synonymous codon usage. When more and more completely sequenced thermophile/mesophile genome pairs of the same genus will be available, it would be interesting to compare those pairs to ascertain the influence of growth temperature on the codon usage of genes having different functional categories.

Acknowledgements: The authors are thankful to the Department of Biotechnology, Government of India, for financial help.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2006.06.014.

References

[1] Grantham, R., Gautier, C., Gouy, M. and Mercier, R. (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, 49r–62r.
[2] Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, 43r–74r.
[3] Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.
[4] Bulmer, M. (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.
[5] Akashi, H. and Eyre-Walker, A. (1998) Translational selection and molecular evolution. Curr. Opin. Genet. Dev. 8, 688–693.
[6] Akashi, H. (2001) Gene expression and molecular evolution. Curr. Opin. Genet. Dev. 11, 660–666.
[7] Gupta, S.K. and Ghosh, T.C. (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene 273, 63–70.
[8] Gupta, S.K., Bhattacharyya, T.K. and Ghosh, T.C. (2004) Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. J. Biomol. Struct. Dyn. 21, 527–536.
[9] Basak, S., Banerjee, T., Gupta, S.K. and Ghosh, T.C. (2004) Investigation on the causes of codon and amino acid usages variation between thermophilic *Aquifex aeolicus* and mesophilic *Bacillus subtilis*. J. Biomol. Struct. Dyn. 22, 205–214.
[10] Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. 151, 389–409.
[11] Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–14.
[12] Sharp, P.M., Tuohy, T. and Mosurski, K. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids. Res. 14, 5125–5143.

[13] Sharp, P.M. and Li, W.-H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 222–230.
[14] Chiapello, H., Lisacek, F., Caboche, M. and Henaut, A. (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209, 1–38.
[15] Liu, Q., Dou, S., Ji, Z. and Xue, Q. (2005) Synonymos codon usage and gene function are strongly related in *Oryza sativa*. BioSystems 80, 123–131.
[16] Lynn, D.J., Singer, G.A.C. and Hickey, D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res. 30, 4272–4277.
[17] McDonald, J.H., Grasso, A.M. and Rejto, L.K. (1999) Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. Mol. Biol. Evol. 16, 1785–1790.
[18] Haney, P.J., Badger, J.H., Buldak, G.L., Reich, C.I. and Woese, C.R. (1999) Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc. Natl. Acad. Sci. USA 96, 3578–3583.
[19] Hendrickson, E.L. et al. (2004) Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. J. Bacteriol. 186, 6956–6969.
[20] Bultz, C.J. et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273, 1058–1073.
[21] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[22] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29, 22–28.
[23] Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.
[24] Wright, F. (1990) The 'effective number of codons' used in a gene. Gene 87, 23–29.
[25] Banerjee, T., Gupta, S.K. and Ghosh, T.C. (2005) Towards a resolution on the inherent methodological weakness of the "effective number of codons used by a gene. Biochem. Biophys. Res. Commun. 330, 1015–1018.
[26] Greenacre, M.J. (1984) Theory and Applications of Correspondence Analysis, Academic Press, London, UK.
[27] Katz, L. and Burge, C.B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. 13, 2042–2051.
[28] Eyre-Walker, A. and Bulmer, M. (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. 21, 4599–4603.